The Amount of DNA Polymorphism Maintained in a Finite Population When the Neutral Mutation Rate Varies Among Sites

Fumio Tajima

Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113, Japan

Manuscript received December 28, 1995
Accepted for publication April 15, 1996

ABSTRACT

The expectations of the average number of nucleotide differences per site (π), the proportion of segregating site (s), the minimum number of mutations per site (s^*) and some other quantities were derived under the finite site models with and without rate variation among sites, where the finite site models include Jukes and Cantor's model, the equal-input model and Kimura's model. As a model of rate variation, the gamma distribution was used. The results indicate that if distribution parameter α is small, the effect of rate variation on these quantities are substantial, so that the estimates of θ based on the infinite site model are substantially underestimated, where $\theta = 4Nv$, N is the effective population size and v is the mutation rate per site per generation. New methods for estimating θ are also presented, which are based on the finite site models with and without rate variation. Using these methods, underestimation can be corrected.

THE amount of DNA polymorphism maintained in a population can be estimated from the average number of pairwise nucleotide differences per site (π) or from the proportion of segregating site (s) among a sample of DNA sequences. When the population is panmictic and at equilibrium and when mutations are selectively neutral, KIMURA (1969) and WATTERSON (1975) showed by using the infinite site model that the expectations of π and s are given by

$$E(\pi) = \theta \quad \text{and} \quad E(s) = a_1(n)\theta, \tag{1}$$

where $\theta = 4Nv$, N is the effective population size, v is the neutral mutation rate per site per generation, and

$$a_1(n) = \sum_{i=1}^{n-1} \frac{1}{i}. \tag{2}$$

These equations suggest that θ can be estimated by

$$\hat{\theta} = \pi, \tag{3a}$$

$$\hat{\theta} = s/a_1(n). \tag{3b}$$

The infinite site model assumes that at most one mutation occurs in each site. However, this is not the case since more than one mutation can occur in each site. This means that (3) might give a biased estimate of θ.

Recently, ROGERS (1992) and BERTORELLE and SLATKIN (1995) have examined the properties of π and s by assuming that each site can have two possible states and that the neutral mutation rate varies among sites, and obtained the following conclusion: When θ is not small and when the neutral mutation rate varies among sites, (3) gives an underestimate of θ although the degree of underestimation depends on the value of θ and the strength of rate variation. The assumption that each site can have only two possible states, however, is not correct since each site can have four nucleotides (A, G, C, and T). In this paper I will present the properties of π and s by using finite site models with four possible nucleotides per site, which are more general than the model used by ROGERS (1992) and BERTORELLE and SLATKIN (1995). If a site has three (or four) nucleotides, we know that at least two (or three) mutations occurred in this site. Thus, the minimum number of mutations is defined as the number of nucleotides minus one, and I will present the property of the minimum number of mutations per site (s^*). I will also develop new methods for estimating θ, based on the finite site models with and without rate variation.

Corresponding author: F. Tajima, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan.

THEORY

In this paper we assume that the population is panmictic and at equilibrium and that the population size (N) is constant. We also assume that mutations are selectively neutral (KIMURA 1968a, 1983).

Jukes and Cantor's model of mutation without rate variation: First, we consider the case where the pattern of mutation follows JUKES and CANTOR's (1969) model. Namely, we assume that in each site the neutral mutation rate is the same among different nucleotides. Denote the mutation rate per nucleotide site per generation by v and the relative frequency of nucleotide i in a particular site in the population by x_i (A, G, C, and T are denoted by nucleotides 1, 2, 3, and 4, respectively). Then, the probability distribution of x_i is given by

TABLE 1

$E(\pi)$, $E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ under Jukes and Cantor's model of mutation without rate variation

		$E(s)/a_1(n)$				$E(s^*)/a_1(n)$			
θ	$E(\pi)$	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 20$	$n = 50$	$n = 100$	$n = 200$
0.005	0.0050	0.0050	0.0049	0.0049	0.0049	0.0050	0.0050	0.0050	0.0050
0.01	0.0099	0.0098	0.0098	0.0097	0.0097	0.0099	0.0099	0.0099	0.0099
0.02	0.0195	0.0192	0.0190	0.0189	0.0188	0.0196	0.0195	0.0195	0.0195
0.05	0.0469	0.0451	0.0442	0.0436	0.0429	0.0474	0.0472	0.0471	0.0469
0.1	0.0882	0.0817	0.0788	0.0766	0.0744	0.0900	0.0895	0.0890	0.0883

$E(\pi)$, $E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ were obtained from Equations 8, 6 and 15, respectively.

$$\phi(x_i) = \frac{\Gamma(4\theta/3)}{\Gamma(\theta)\Gamma(\theta/3)}(1 - x_i)^{\theta-1}x_i^{\theta/3-1} \quad (4)$$

(KIMURA 1968b). Suppose now that n DNA sequences are randomly sampled from the population. Using the Ewens sampling theory (EWENS 1972), the probability (p_i) that a particular site in the sample is exclusively occupied by nucleotide i is given by

$$p_i = \int_0^1 \phi(x_i)\, x_i^n dx = \frac{\Gamma(4\theta/3)\Gamma(\theta/3 + n)}{\Gamma(4\theta/3 + n)\Gamma(\theta/3)}. \quad (5)$$

Then, the expectation of s is given by

$$E(s) = 1 - \sum_{i=1}^{4} p_i = 1 - 4p_i, \quad (6)$$

which is approximately given by

$$E(s) \approx a_1(n)\theta \exp\{-c_1(n)\theta\}, \quad (6a)$$

$$E(s) \approx \frac{a_1(n)\theta}{1 + c_1(n)\theta}, \quad (6b)$$

where $c_1(n)$ is given by $c_1(n) = 4a_1(n)/3 - 5a_2(n)/\{3a_1(n)\}$, $a_1(n)$ is given by (2) and $a_2(n)$ is given by

$$a_2(n) = \left[\{a_1(n)\}^2 - \sum_{i=1}^{n-1}\frac{1}{i^2}\right]\Big/ 2. \quad (7)$$

Note that $a_1(n) = -S_n^{(2)}/S_n^{(1)}$ and $a_2(n) = S_n^{(3)}/S_n^{(1)}$, where $S_n^{(i)}$ is the Stirling number of the first kind.

The expectation of π can be obtained by substituting $n = 2$ into (6), and we have

$$E(\pi) = \frac{\theta}{1 + 4\theta/3}, \quad (8)$$

which agrees with Equation 15 of TAJIMA (1983). Numerical examples of $E(\pi)$ and $E(s)/a_1(n)$ are given in Table 1 since π and $s/a_1(n)$ are used to estimate θ under the infinite site model. This table shows that π and $s/a_1(n)$ give underestimates of θ, that the underestimation is substantial only when θ is large, and that the degree of underestimation is larger for $s/a_1(n)$ than for π. It is also noted that the bias of $S/a_1(n)$ increases as the sample size (n) increases. This is because, as n increases, the probability that more than one mutation occurred in each site among a sample of n sequences also increases. Equations 8 and 6b suggest that θ can be estimated by

$$\hat{\theta} = \frac{\pi}{1 - 4\pi/3}, \quad (9)$$

$$\hat{\theta} = \frac{s}{a_1(n) - c_1(n)s}. \quad (10)$$

We can estimate θ from the minimum number of mutations per site (s^*). Let p_{ij} be the probability that a particular site is exclusively occupied by nucleotides i and/or j, and p_{ijk} be the probability that a particular site is exclusively occupied by nucleotides i, j and/or k. Using the same method as we obtained (5), we obtain

$$p_{ij} = \frac{\Gamma(4\theta/3)\Gamma(2\theta/3 + n)}{\Gamma(4\theta/3 + n)\Gamma(2\theta/3)}, \quad (11)$$

$$p_{ijk} = \frac{\Gamma(4\theta/3)\Gamma(\theta + n)}{\Gamma(4\theta/3 + n)\Gamma(\theta)}. \quad (12)$$

Denote the probability that a particular site is occupied by i types of nucleotide in a sample of n DNA sequences by q_i and the estimate of q_i by $\hat{q}_i$. For example, if we have the following five DNA sequences with a length of 20 nucleotides where dash (-) indicates the same nucleotide as in the first sequence listed, we have $\hat{q}_1 = {}^{12}/_{20}$, $\hat{q}_2 = {}^{5}/_{20}$, $\hat{q}_3 = {}^{2}/_{20}$ and $\hat{q}_4 = {}^{1}/_{20}$.

```
AGCCTACTTAATCGTAGGAC
---A---C---------C--
G--A---C-----C---C--
G--A-G-C-----C-G-A--
C-TA-G-C-----T-G-T--
```

It is clear from the definitions that we have

$$q_1 = 4p_i, \quad (13a)$$

$$q_2 = 6p_{ij} - 12p_i, \quad (13b)$$

$$q_3 = 4p_{ijk} - 12p_{ij} + 12p_i, \quad (13c)$$

$$q_4 = 1 - 4p_{ijk} + 6p_{ij} - 4p_i. \quad (13d)$$

When a site is occupied by i types of nucleotide in the sample, it is certain that at least $i - 1$ mutations occurred in this site. Therefore, the minimum number of mutations per site can be estimated by

$$s^* = \hat{q}_2 + 2\hat{q}_3 + 3\hat{q}_4. \quad (14)$$

In the above example, we have $s^* = {}^5/_{20} + 2 \times {}^2/_{20} + 3 \times {}^1/_{20} = {}^{12}/_{20} = 0.6$. It should be noted that s^* is the same as s in the infinite site model, but usually different in the finite site model since $s = \hat{q}_2 + \hat{q}_3 + \hat{q}_4$. From (13) the expectation of s^* can be given by

$$E(s^*) = q_2 + 2q_3 + 3q_4 = 3 - 4p_{ijk}, \quad (15)$$

which is approximately given by

$$E(s^*) \approx \frac{a_1(n)\theta}{1 + c_2(n)\theta}, \quad (15a)$$

where $c_2(n) = 4a_1(n)/3 - 7a_2(n)/\{3a_1(n)\}$. Numerical examples of $E(s^*)/a_1(n)$ are also shown in Table 1, which indicate that $s^*/a_1(n)$ gives slightly better estimates of θ than does $s/a_1(n)$. Equation 15a suggests that θ can be estimated by

$$\hat{\theta} = \frac{s^*}{a_1(n) - c_2(n)s^*}. \quad (16)$$

Under the infinite site model, the expected number of nucleotides whose frequency is i/n per site is given by

$$f_n(i) = \theta\left(\frac{1}{i} + \frac{1}{n-i}\right) \quad (17)$$

(TAJIMA 1989a), which can be used to identify the pattern of DNA polymorphism. Under the present model, it is given by

$$f_n(i) = 4\int_0^1 \frac{n!}{i!(n-i)!} x^i(1-x)^{n-i}\phi(x)\,dx$$

$$= \frac{4\Gamma(n+1)\Gamma(4\theta/3)\Gamma(\theta/3+i)\Gamma(\theta+n-i)}{\Gamma(i+1)\Gamma(n-i+1)\Gamma(4\theta/3+n)\Gamma(\theta/3)\Gamma(\theta)}, \quad (18)$$

which can be approximately given by

$$f_n(i) \approx \frac{\theta\{1/i + 1/(n-i)\}}{1 + \{4a_1(n)/3 - a_1(n-i) - a_1(i)/3\}\theta}. \quad (18a)$$

A comparison between (17) and (18) indicates that $f_n(i)$ under the present model is not very different from that of the infinite site model even if θ is quite large except when i/n is close to 1.

Jukes and Cantor's model of mutation with rate variation: It has been shown by several authors (*e.g.*, GOLDING 1983; WAKELEY 1993) that neutral mutation rates are approximately gamma distributed among sites.

TABLE 2

$E(\pi)$, $E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ under Jukes and Cantor's model of mutation with rate variation

$E(\theta)$	$E(\pi)$	$E(s)/a_1(n)$	$E(s^*)/a_1(n)$
		(A) $\alpha = 0.1$	
0.005	0.0047	0.0043	0.0047
0.01	0.0087	0.0076	0.0088
0.02	0.0155	0.0123	0.0157
		(B) $\alpha = 0.2$	
0.005	0.0048	0.0046	0.0048
0.01	0.0093	0.0085	0.0093
0.02	0.0172	0.0149	0.0174
		(C) $\alpha = 0.5$	
0.005	0.0049	0.0048	0.0049
0.01	0.0096	0.0092	0.0096
0.02	0.0185	0.0171	0.0186
		(D) $\alpha = 1$	
0.005	0.0049	0.0049	0.0049
0.01	0.0097	0.0095	0.0098
0.02	0.0190	0.0180	0.0191

$E(\pi)$, $E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ were obtained from formulas 22, 21 and 23, respectively, where $n = 100$ is assumed.

Here, we assume that θ follows the following gamma distribution:

$$g(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\theta}\theta^{\alpha-1}, \quad (19)$$

where $\alpha = \{E(\theta)\}^2/V(\theta)$, $\beta = \alpha/E(\theta)$ and $E(\theta)$ is the expectation of θ, *i.e.*, $E(\theta) = \int \theta g(\theta)\,d\theta$. Note that the smaller α is, more the mutation rate varies among sites. Then, using (6), the expectation of the proportion of segregating site is given by

$$E(s) = \int_0^\infty (1 - 4p_i)\,g(\theta)\,d\theta. \quad (20)$$

Using (6a), $E(s)$ is approximately given by

$$E(s) \approx \frac{a_1(n)E(\theta)}{1 + c_1(n)(\alpha+1)E(\theta)/\alpha}. \quad (21)$$

Substituting $n = 2$ into (21), we have

$$E(\pi) \approx \frac{E(\theta)}{1 + 4(\alpha+1)E(\theta)/(3\alpha)}. \quad (22)$$

In the same way, the expectation of the minimum number of mutations per site can be obtained, which is approximately given by

$$E(s^*) \approx \frac{a_1(n)E(\theta)}{1 + c_2(n)(\alpha+1)E(\theta)/\alpha}. \quad (23)$$

Numerical examples are shown in Table 2, where $E(\pi)$,

$E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ are given for $n = 100$. We can see from this table that the effect of rate variation is substantial when α is small ($\alpha < 1$) even if $E(\theta)$ is as small as 0.005. The effect is stronger on $E(s)$ than on $E(\pi)$ and $E(s^*)$. Formulas (22), (21) and (23) suggest that we can estimate θ by

$$\hat{\theta} = \frac{\pi}{1 - 4(\alpha + 1)\pi/(3\alpha)}, \quad (24)$$

$$\hat{\theta} = \frac{s}{a_1(n) - c_1(n)(\alpha + 1)s/\alpha}, \quad (25)$$

$$\hat{\theta} = \frac{s^*}{a_1(n) - c_2(n)(\alpha + 1)s^*/\alpha}, \quad (26)$$

if α is known. Although we can also estimate α from these formulas, the accuracy of the estimate might be very low since the variances of s and π are too large to estimate an additional parameter (WATTERSON 1975; TAJIMA 1983).

Equal-input model of mutation without rate variation: Here, we assume that the mutation rate to nucleotide i from any of the other three nucleotides is the same (FELSENSTEIN 1981; TAJIMA and NEI 1982). Namely, when v_{ji} is the mutation rate per generation from nucleotide j to nucleotide i, we assume $v_{ji} = v_i$ for $j \neq i$. Under this model, the mutation rate per generation from nucleotide i to any of the other three nucleotides is $\Sigma_{j\neq i} v_j$ and the expected frequency of nucleotide i is given by

$$y_i = v_i \Big/ \sum_{j=1}^{4} v_j \quad (27)$$

(TAJIMA and NEI 1984). Then, the expected mutation rate per site per generation can be given by

$$v = \sum_{i=1}^{4} y_i \sum_{j\neq i} v_{ij} = h_1 \sum_{i=1}^{4} v_i, \quad (28)$$

where $h_1 = 1 - \Sigma_{i=1}^{4} y_i^2$. Following KIMURA (1968b), the probability distribution of the frequency of nucleotide i, x_i, is given by

$$\phi(x_i) = \frac{\Gamma(\theta/h_1)}{\Gamma[(1 - y_i)\theta/h_1]\Gamma(y_i\theta/h_1)} \times (1 - x_i)^{(1-y_i)\theta/h_1 - 1} x_i^{y_i\theta/h_1 - 1}, \quad (29)$$

since $4Nv_i = y_i\theta/h_1$ and $4N\Sigma_{j\neq i} v_j = (1 - y_i)\theta/h_1$, where $\theta = 4Nv$. Using the Ewens sampling theory (EWENS 1972), the probability (p_i) that a particular site in the sample is exclusively occupied by nucleotide i is given by

$$p_i = \int_0^1 \phi(x_i)\, x_i^n\, dx = \frac{\Gamma(\theta/h_1)\Gamma(y_i\theta/h_1 + n)}{\Gamma(\theta/h_1 + n)\Gamma(y_i\theta/h_1)}. \quad (30)$$

Then, the expectation of s is given by

TABLE 3

$E(\pi)$, $E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ under the equal-input model of mutation without rate variation

θ	$E(\pi)$	$E(s)/a_1(n)$	$E(s^*)/a_1(n)$
(A) Moderately unequal nucleotide frequencies			
0.005	0.0050	0.0049	0.0050
0.01	0.0099	0.0097	0.0099
0.02	0.0194	0.0189	0.0194
0.05	0.0467	0.0434	0.0466
0.1	0.0875	0.0761	0.0873
(B) Extremely unequal nucleotide frequencies			
0.005	0.0049	0.0049	0.0049
0.01	0.0098	0.0096	0.0097
0.02	0.0192	0.0186	0.0190
0.05	0.0453	0.0421	0.0442
0.1	0.0830	0.0723	0.0793

$E(\pi)$, $E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ were obtained from Equations 32, 31 and 35, respectively, where $n = 100$ is assumed. Moderately unequal nucleotide frequencies assume $y_1 = 0.1$, $y_2 = 0.3$, $y_3 = 0.4$ and $y_4 = 0.2$, and extremely unequal nucleotide frequencies assume $y_1 = 0.025$, $y_2 = 0.225$, $y_3 = 0.675$ and $y_4 = 0.075$.

$$E(s) = 1 - \sum_{i=1}^{4} p_i, \quad (31)$$

which is approximately given by

$$E(s) \approx \frac{a_1(n)\theta}{1 + \left\{\frac{a_1(n)}{h_1} - \frac{a_2(n)h_2}{a_1(n)h_1^2}\right\}\theta}, \quad (31a)$$

where $h_2 = 1 - \Sigma_{i=1}^{4} y_i^3$. On the other hand, the expectation of π is given by

$$E(\pi) = \frac{\theta}{1 + \theta/h_1}. \quad (32)$$

In the same way as before, p_{ij} and p_{ijk} can be obtained, which are given by

$$p_{ij} = \frac{\Gamma(\theta/h_1)\Gamma[(y_i + y_j)\theta/h_1 + n]}{\Gamma(\theta/h_1 + n)\Gamma[(y_i + y_j)\theta/h_1]}, \quad (33)$$

$$p_{ijk} = \frac{\Gamma(\theta/h_1)\Gamma[(y_i + y_j + y_k)\theta/h_1 + n]}{\Gamma(\theta/h_1 + n)\Gamma[(y_i + y_j + y_k)\theta/h_1]}. \quad (34)$$

Then, the expectation of s^* is given by

$$E(s^*) = 3 - \sum_{i<j<k} p_{ijk}, \quad (35)$$

which is approximately given by

$$E(s^*) \approx \frac{a_1(n)\theta}{1 + \left\{\frac{a_1(n)}{h_1} - \frac{a_2(n)(3h_1 - h_2)}{a_1(n)h_1^2}\right\}\theta}. \quad (35a)$$

Numerical examples are shown in Table 3, where $n =$

100 is assumed. A comparison between Tables 1 and 3 shows that $E(\pi)$ and $E(s)$ under this model are not very different from those under Jukes and Cantor's model even when the nucleotide frequency deviates substantially from equality, whereas $E(s^*)$ under this model is quite different from that under Jukes and Cantor's model when the nucleotide frequency deviates substantially from equality. Using (31a), (32) and (35a), we can estimate θ by

$$\hat{\theta} = \frac{\pi}{1 - \pi/\hat{h}_1}, \tag{36}$$

$$\hat{\theta} = \frac{s}{a_1(n) - \left\{\frac{a_1(n)}{\hat{h}_1} - \frac{a_2(n)\hat{h}_2}{a_1(n)\hat{h}_1^2}\right\} s}, \tag{37}$$

$$\hat{\theta} = \frac{s^*}{a_1(n) - \left\{\frac{a_1(n)}{\hat{h}_1} - \frac{a_2(n)(3\hat{h}_1 - \hat{h}_2)}{a_1(n)\hat{h}_1^2}\right\} s^*}. \tag{38}$$

In these equations, $\hat{h}_1$ and $\hat{h}_2$ can be estimated by $\hat{h}_1 = 1 - \Sigma_{i=1}^4 \hat{y}_i^2$ and $\hat{h}_2 = 1 - \Sigma_{i=1}^4 \hat{y}_i^3$, where $\hat{y}_i$ is the observed frequency of nucleotide i.

Under this model the expected number of nucleotides whose frequency is i/n per site is given by

$$f_n(i) = \sum_{j=1}^{4} \int_0^1 \frac{n!}{i!(n-i)!} x_j^i (1 - x_j)^{n-i} \phi(x_j)\, dx_j$$
$$= \frac{\Gamma(n+1)\Gamma(\theta/h_1)}{\Gamma(i+1)\Gamma(n-i+1)\Gamma(\theta/h_1 + n)} \times \sum_{j=1}^{4} \frac{\Gamma(y_j\theta/h_1 + i)\Gamma[(1-y_j)\theta/h_1 + n - i]}{\Gamma(y_j\theta/h_1)\Gamma[(1-y_j)\theta/h_1]}. \tag{39}$$

Equal-input model of mutation with rate variation: When θ follows the gamma distribution (19), the expectations of π, s and s^* are approximately given by

$$E(\pi) \approx \frac{E(\theta)}{1 + (\alpha+1)E(\theta)/(h_1\alpha)}, \tag{40}$$

$$E(s) \approx \frac{a_1(n)E(\theta)}{1 + \left\{\frac{a_1(n)}{h_1} - \frac{a_2(n)h_2}{a_1(n)h_1^2}\right\}\frac{\alpha+1}{\alpha}E(\theta)}, \tag{41}$$

$$E(s^*) \approx \frac{a_1(n)E(\theta)}{1 + \left\{\frac{a_1(n)}{h_1} - \frac{a_2(n)(3h_1 - h_2)}{a_1(n)h_1^2}\right\}\frac{\alpha+1}{\alpha}E(\theta)}. \tag{42}$$

Numerical examples are shown in Table 4. It can be seen from this table that the effect of rate variation is substantial when α is small ($\alpha < 1$) even if $E(\theta)$ is as small as 0.005 and that the effect is stronger on $E(s)$ than on $E(\pi)$. The above formulas suggest that, if α is known, θ can be estimated by

$$\hat{\theta} = \frac{\pi}{1 - (\alpha+1)\pi/(\hat{h}_1\alpha)}, \tag{43}$$

$$\hat{\theta} = \frac{s}{a_1(n) - \left\{\frac{a_1(n)}{\hat{h}_1} - \frac{a_2(n)\hat{h}_2}{a_1(n)\hat{h}_1^2}\right\}\frac{\alpha+1}{\alpha}s}, \tag{44}$$

$$\hat{\theta} = \frac{s^*}{a_1(n) - \left\{\frac{a_1(n)}{\hat{h}_1} - \frac{a_2(n)(3\hat{h}_1 - \hat{h}_2)}{a_1(n)\hat{h}_1^2}\right\}\frac{\alpha+1}{\alpha}s^*}. \tag{45}$$

Kimura's model of mutation without rate variation: We now assume that the rate of transitional mutation is different from that of transversional mutation (KIMURA 1980). Namely, we assume $v_{12} = v_{21} = v_{34} = v_{43} = \sigma v$ and $v_{13} = v_{14} = v_{23} = v_{24} = v_{31} = v_{32} = v_{41} = v_{42} = (\omega/2)v$ and $\sigma + \omega = 1$, where v_{ij} is the mutation rate per site per generation from nucleotide i to nucleotide j, v is the total mutation rate per site per generation, σ is the proportion of transitional mutation, and ω is the proportion of transversional mutation. Unlike the previous models, under this model we cannot obtain the expectations of s and s^* since we cannot obtain p_i. We can, however, obtain p_{ij} as follows. The mutation rate from A or G to C or T is $2\omega v$, and the mutation rate from C or T to A or G is also $2\omega v$. Therefore, $p_{12}(=p_{34})$ is given by

$$p_{12} = \frac{\Gamma(2\omega\theta)\Gamma(\omega\theta + n)}{\Gamma(2\omega\theta + n)\Gamma(\omega\theta)}. \tag{46}$$

Since the mutation rate from A or C to G or T and that from G or T to A or C are $(\sigma + \omega/2)v$, p_{13} $(=p_{14} = p_{23} = p_{24})$ is given by

$$p_{13} = \frac{\Gamma[(2\sigma+\omega)\theta]\Gamma[(\sigma+\omega/2)\theta + n]}{G[(2\sigma+\omega)\theta + n]\Gamma[(\sigma+\omega/2)\theta]}. \tag{47}$$

Denote the probability that a particular site is occupied by (A and G) or (C and T) in a sample of n DNA sequences by q_s, and the probability of having (A and C), (A and T), (G and C) or (G and T) by q_v. In other words, q_2 is divided into the transitional part (q_s) and the transversional part (q_v). In the previous example, the estimates of q_s and q_v are $^4/_{20}$ and $^1/_{20}$, respectively. If we denote q_a and q_b by

$$q_a = q_v + q_3 + q_4 \text{ and } q_b = q_s + q_v/2 + q_3 + q_4, \tag{48}$$

q_a and q_b are given by

TABLE 4

$E(\pi)$, $E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ under the equal-input model of mutation with rate variation

	Moderately unequal nucleotide frequencies[a]			Extremely unequal nucleotide frequencies[a]		
$E(\theta)$	$E(\pi)$	$E(s)/a_1(n)$	$E(s^*)/a_1(n)$	$E(\pi)$	$E(s)/a_1(n)$	$E(s^*)/a_1(n)$
			(A) $\alpha = 0.1$			
0.005	0.0046	0.0043	0.0046	0.0045	0.0042	0.0044
0.01	0.0086	0.0076	0.0086	0.0082	0.0071	0.0078
0.02	0.0152	0.0122	0.0152	0.0138	0.0111	0.0127
			(B) $\alpha = 0.2$			
0.005	0.0048	0.0046	0.0048	0.0047	0.0045	0.0046
0.01	0.0092	0.0085	0.0092	0.0089	0.0082	0.0086
0.02	0.0171	0.0148	0.0170	0.0160	0.0139	0.0152
			(C) $\alpha = 0.5$			
0.005	0.0049	0.0048	0.0049	0.0049	0.0047	0.0048
0.01	0.0096	0.0092	0.0096	0.0094	0.0090	0.0093
0.02	0.0184	0.0170	0.0184	0.0178	0.0164	0.0173
			(D) $\alpha = 1$			
0.005	0.0049	0.0049	0.0049	0.0049	0.0048	0.0049
0.01	0.0097	0.0094	0.0097	0.0096	0.0093	0.0095
0.02	0.0189	0.0179	0.0189	0.0185	0.0174	0.0181

$E(\pi)$, $E(s)/a_1(n)$ and $E(s^*)/a_1(n)$ were obtained from formulas 40, 41 and 42, respectively, where $n = 100$ is assumed.

[a] See Table 3.

$$q_a = 1 - 2p_{12} \quad \text{and} \quad q_b = 1 - 2p_{13}, \tag{49}$$

which are approximately given by

$$q_a \approx \frac{a_1(n)\omega\theta}{1 + c_3(n)\omega\theta} \quad \text{and} \quad q_b \approx \frac{a_1(n)(\sigma + \omega/2)\theta}{1 + c_3(n)(\sigma + \omega/2)\theta}, \tag{50}$$

where $c_3(n) = 2a_1(n) - 3a_2(n)/a_1(n)$. These formulas suggest that θ, ω and σ can be estimated by

$$\hat{\theta} = \frac{\hat{q}_a}{2\{a_1(n) - c_3(n)\hat{q}_a\}} + \frac{\hat{q}_b}{a_1(n) - c_3(n)\hat{q}_b}, \tag{51a}$$

$$\hat{\omega} = \frac{\hat{q}_a}{\{a_1(n) - c_3(n)\hat{q}_a\}\hat{\theta}} \quad \text{and} \quad \hat{\sigma} = 1 - \hat{\omega}. \tag{51b}$$

Let us now denote the average numbers of pairwise transitional and transversional differences per nucleotide site by π_s and π_v, respectively. When $n = 2$, we have $q_a = E(\pi_v)$ and $q_b = E(\pi_s) + E(\pi_v)/2$ since $E(\pi_s) = q_s$ and $E(\pi_v) = q_v$. From (49), we have

$$E(\pi_v) = \frac{\omega\theta}{1 + 2\omega\theta}, \tag{52a}$$

$$E(\pi_s) + E(\pi_v)/2 = \frac{(\sigma + \omega/2)\theta}{1 + (2\sigma + \omega)\theta}. \tag{52b}$$

Therefore, the expectation of π can be given by

$$E(\pi) = E(\pi_s) + E(\pi_v) = \frac{\{1 + 3(1 - \omega/2)\omega\theta\}\theta}{(1 + 2\omega\theta)\{1 + (2 - \omega)\theta\}}. \tag{53}$$

It is noted that $E(\pi) = \theta/(1 + 4\theta/3)$ if there is no transition bias (*i.e.*, $\omega = {}^2/_3$). Numerical examples are shown in Table 5, which indicate that π underestimates θ when θ is very large and that the amount of underestimation is the slightest when there is no transition bias. Equations 52a and 52b indicate that we can estimate θ, ω and σ by

$$\hat{\theta} = \frac{\pi_s + \pi_v/2}{1 - (2\pi_s + \pi_v)} + \frac{\pi_v}{2(1 - 2\pi_v)}, \tag{54a}$$

$$\hat{\omega} = \frac{\pi_v}{(1 - 2\pi_v)\hat{\theta}} \quad \text{and} \quad \hat{\sigma} = 1 - \hat{\omega}. \tag{54b}$$

Kimura's model of mutation with rate variation: We again assume that θ follows the gamma distribution (19). In the same way as before, we can obtain the expectation of q_a and q_b, which are approximately given by

$$E(q_a) \approx \frac{a_1(n)\omega E(\theta)}{1 + c_3(n)(\alpha + 1)\omega E(\theta)/\alpha}, \tag{55a}$$

TABLE 5

$E(\pi)$ under Kimura's model of mutation without rate variation

		θ				
σ	ω	0.005	0.01	0.02	0.05	0.1
0.0	1.0	0.0050	0.0099	0.0194	0.0465	0.0871
0.1	0.9	0.0050	0.0099	0.0195	0.0467	0.0877
0.2	0.8	0.0050	0.0099	0.0195	0.0468	0.0881
0.3	0.7	0.0050	0.0099	0.0195	0.0469	0.0882
0.4	0.6	0.0050	0.0099	0.0195	0.0469	0.0882
0.5	0.5	0.0050	0.0099	0.0195	0.0468	0.0879
0.6	0.4	0.0050	0.0099	0.0194	0.0467	0.0875
0.7	0.3	0.0050	0.0098	0.0194	0.0465	0.0868
0.8	0.2	0.0050	0.0098	0.0194	0.0462	0.0859
0.9	0.1	0.0050	0.0098	0.0193	0.0459	0.0847
1.0	0.0	0.0050	0.0098	0.0192	0.0455	0.0833

$E(\pi)$ was obtained from Equation 53.

$$E(q_b) \approx \frac{a_1(n)(\sigma + \omega/2)E(\theta)}{1 + c_3(n)(\alpha + 1)(\sigma + \omega/2)E(\theta)/\alpha}. \quad (55b)$$

Then, if α is known, θ, ω and σ can be estimated by

$$\hat{\theta} = \frac{\hat{q}_a}{2\{a_1(n) - c_3(n)(\alpha + 1)\hat{q}_a/\alpha\}} + \frac{\hat{q}_b}{a_1(n) - c_3(n)(\alpha + 1)\hat{q}_b/\alpha}, \quad (56a)$$

$$\hat{\omega} = \frac{\hat{q}_a}{\{a_1(n) - c_3(n)(\alpha + 1)\hat{q}_a/\alpha\}\hat{\theta}} \quad \text{and} \quad \hat{\sigma} = 1 - \hat{\omega}. \quad (56b)$$

The expectation of π can be obtained from $E(\pi) = E(\pi_s) + E(\pi_v)$. Since $E(\pi_v)$ and $E(\pi_s) + E(\pi_v)/2$ are approximately given by

$$E(\pi_v) \approx \frac{\omega\theta}{1 + 2(\alpha + 1)\omega\theta/\alpha}, \quad (57a)$$

$$E(\pi_s) + E(\pi_v)/2 \approx \frac{(\sigma + \omega/2)\theta}{1 + (\alpha + 1)(2\sigma + \omega)\theta/\alpha}, \quad (57b)$$

the expectation of π is approximately given by

$$E(\pi) \approx \frac{\{1 + 3(\alpha + 1)(1 - \omega/2)\omega\theta/\alpha\}\theta}{\{1 + 2(\alpha + 1)\omega\theta/\alpha\}\{1 + (\alpha + 1)(2 - \omega)\theta/\alpha\}}. \quad (58)$$

Numerical examples are shown in Table 6, which indicate that the effect of rate variation on π is substantial even when $E(\theta)$ is as small as 0.005 and that the effect increases as the transition/transversion bias increases from $\sigma = {}^1/_3$. Formulas 57a and 57b suggest that if α is known, θ, ω and σ can be estimated by

$$\hat{\theta} = \frac{\pi_s + \pi_v/2}{1 - (\alpha + 1)(2\pi_s + \pi_v)/\alpha} + \frac{\pi_v}{2(1 - 2(\alpha + 1)\pi_v/\alpha)}, \quad (59a)$$

$$\hat{\omega} = \frac{\pi_v}{(1 - 2(\alpha + 1)\pi_v/\alpha)\hat{\theta}} \quad \text{and} \quad \hat{\sigma} = 1 - \hat{\omega}. \quad (59b)$$

DISCUSSION

The infinite site model is one of the fundamental models for molecular population genetics. In fact, a large number of important theoretical studies were based on this model. This model, however, is not always applicable. As shown in this paper, when the rate of mutation varies substantially among sites, the amount of DNA polymorphism estimated by using this model might be underestimated. For example, the amount of underestimation might be substantial in the case of the control region of human mitochondrial DNA. HORAI and coworkers examined the 482-bp sequences in the control region from Africans, Europeans, Asians and Native Americans (HORAI and HAYASAKA 1990; HORAI *et al.* 1991, 1993). I have analyzed the 250-bp sequences in hypervariable region 1, since an estimate of α is available in this region, *i.e.*, $\hat{\alpha} = 0.47$ (WAKELEY 1993). Table 7 shows the estimates of θ, σ and ω obtained from the infinite site model, Jukes and Cantor's model, the equal-input model and Kimura's model by using $n = 193$, $\hat{q}_2 = {}^{95}/_{250}$ ($\hat{q}_s = {}^{91}/_{250}$ and $\hat{q}_v = {}^4/_{250}$), $\hat{q}_3 = {}^5/_{250}$, $\hat{q}_4 = 0$, $\hat{y}_1 = 0.35$, $\hat{y}_2 = 0.09$, $\hat{y}_3 = 0.37$ and $\hat{y}_4 = 0.19$. It should be noted that in this case $\theta = 2N_f v$, where N_f is the effective number of females, since mitochondrial DNA is haploid and maternally inherited. The results show that the estimate based on the infinite site model is smaller than the estimates based on the other three models. In the cases of Jukes and Cantor's model and the equal-input model, there are discrepancies between

TABLE 6

$E(\pi)$ under Kimura's model of mutation with rate variation

	σ					
$E(\theta)$	0.0	0.2	0.4	0.6	0.8	1.0
			(A) $\alpha = 0.1$			
0.005	0.0046	0.0047	0.0047	0.0046	0.0046	0.0045
0.01	0.0086	0.0087	0.0087	0.0086	0.0085	0.0082
0.02	0.0151	0.0154	0.0155	0.0152	0.0147	0.0139
			(B) $\alpha = 0.2$			
0.005	0.0048	0.0048	0.0048	0.0048	0.0048	0.0047
0.01	0.0092	0.0092	0.0093	0.0092	0.0091	0.0089
0.02	0.0170	0.0172	0.0172	0.0171	0.0167	0.0161
			(C) $\alpha = 0.5$			
0.005	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
0.01	0.0096	0.0096	0.0096	0.0096	0.0095	0.0094
0.02	0.0184	0.0185	0.0185	0.0184	0.0182	0.0179
			(D) $\alpha = 1$			
0.005	0.0049	0.0049	0.0049	0.0049	0.0049	0.0049
0.01	0.0097	0.0097	0.0097	0.0097	0.0097	0.0096
0.02	0.0189	0.0190	0.0190	0.0189	0.0188	0.0185

$E(\pi)$ was obtained from (58).

the estimates obtained from s and s^*, probably because these mutation models do not fit the data. Since the transitional mutation rate is substantially higher than transversional mutation rate in mammalian mitochondrial DNA sequences, Kimura's model might be more appropriate. It is also shown in the table that in the case of Kimura's model the estimate of θ with rate variation is quite different from that without rate variation. Namely the value with rate variation is 2.5 times larger than that without rate variation. $\hat{\theta} \approx 0.2$ may not be unreasonable, since θ is determined by the recent population size more strongly than by the ancient population size (TAJIMA 1989b) and since the human population might have increased (MERRIWETHER *et al.* 1991; ROGERS and HARPENDING 1992).

Recently, several methods for estimating θ, including FU (1994a,b) and KUHNER *et al.* (1995), have been proposed. FU's methods, however, assume that there is no rate variation among nucleotide sites, so that they may give an underestimate of θ when the mutation rate

TABLE 7

Estimates of θ, σ and ω in the 250-bp hypervariable control region of human mitochondrial DNA

Model	Measure used	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\omega}$	Formula
Infinite site model	s	0.0685			3b
Jukes and Cantor's model					
Without rate variation	s	0.0874			10
	s^*	0.0794			16
With rate variation	s	0.2113			25
	s^*	0.1017			26
Equal-input model					
Without rate variation	s	0.0880			37
	s^*	0.0811			38
With rate variation	s	0.2231			44
	s^*	0.1110			45
Kimura's model					
Without rate variation	$\hat{q}_a$, $\hat{q}_b$	0.0897	0.9298	0.0702	51
With rate variation	$\hat{q}_a$, $\hat{q}_b$	0.2283	0.9711	0.0289	56

Data from HORAI *et al.* (1993). In the case of rate variation, $\alpha = 0.47$ was used.

varies among sites. Since we do not know the effect of rate variation on his methods, we must be careful when the rate varies among sites substantially. On the other hand, KUHNER *et al.*'s method can deal with rate variation by using a mutation rate category approach, although this approach has not been extensively tested.

BERTORELLE and SLATKIN (1995) clearly indicate that the test of neutrality proposed by TAJIMA (1989a) might not be appropriate if the mutation rate varies among sites substantially. This is because Tajima's D statistic is based on the infinite site model. FU and LI's (1993) test also has the same problem. The variances of π and s and the covariance between π and s remain to be solved under the finite site models with rate variation. If we know them, we will be able to alleviate this problem.

I thank an anonymous reviewer for many valuable suggestions and comments for improving the presentation. This work was supported in part by a grant-in-aid from the Ministry of Education, Science, Sports and Culture of Japan.

LITERATURE CITED

BERTORELLE, G., and M. SLATKIN, 1995 The number of segregating sites in expanding human populations, with implications of estimates of demographic parameters. Mol. Biol. Evol. **12:** 887–892.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

FU, Y.-X., 1994a A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

FU, Y.-X., 1994b Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics **138:** 1375–1386.

FU, Y.-X., and W.-H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

GOLDING, G. B., 1983 Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. **1:** 125–142.

HORAI, S., and K. HAYASAKA, 1990 Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am. J. Hum. Genet. **46:** 828–842.

HORAI, S., R. KONDO, K. MURAYAMA, S. HAYASHI, H. KOIKE *et al.*, 1991 Phylogenetic affiliation of ancient and contemporary humans inferred from mitochondrial DNA. Phil. Trans. R. Soc. Lond. B **333:** 409–417.

HORAI, S., R. KONDO, Y. NAKAGAWA-HATTORI, S. HAYASHI, S. SONODA *et al.*, 1993 Peopling of the Americas, founded by four major lineages of mitochondrial DNA. Mol. Biol. Evol. **10:** 23–47.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism,* edited by H. N. MUNRO. Academic Press, New York.

KIMURA, M., 1968a Evolutionary rate at the molecular level. Nature **217:** 624–626.

KIMURA, M., 1968b Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11:** 247–269.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

KIMURA, M., 1980 A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

MERRIWETHER, D. A., A. G. CLARK, S. W. BALLINGER, T. G. SCHURR, H. SOODYALL *et al.*, 1991 The structure of human mitochondrial DNA variation. J. Mol. Evol. **33:** 543–555.

ROGERS, A., 1992 Error introduced by the infinite-sites model. Mol. Biol. Evol. **9:** 1181–1184.

ROGERS, A. R., and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

TAJIMA, F., 1989a Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

TAJIMA, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

TAJIMA, F., and M. NEI, 1982 Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. **18:** 115–120.

TAJIMA, F., and M. NEI, 1984 Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1:** 269–285.

WAKELEY, J., 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

WATTERSON, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: G. B. GOLDING